Midterm Examination CS540-1: Introduction to Artificial Intelligence

Oct 26, 2011. (20 questions, 5 points each)

BE SURE TO WRITE YOUR LAST NAME, FIRST NAME, AND EMAIL ON THE ANSWER BOOK.

1. (HAC) Run hierarchical clustering with single linkage on the following 1D dataset, and show the tree: 0, 3, 4, 6, 10, 12.

   ((0 ((34) 6)) (10 12))

2. (k-means) Consider the 1D dataset: $x_i=i$ for i=1 to 25. To select good initial cluster centers for k-means where k=5, let's set c_1=1. Then, select c_j from the unused points in the dataset, so that it is farthest from any already-selected centers $c_1 \ldots c_{j-1}$. Show $c_2 \ldots c_5$.
   c1=1, c2=25, c3=13, c4 and c5=7 and 19 (arbitrary order)

3. (kNN) Consider a 2D dataset where training points are on the integer grid (x_1,x_2), and both dimensions x_1 and x_2 range from 1 to 100 (inclusive). The binary label for point (x_1,x_2) is $(-1)^{(x1+x2)}$. What is the label for test point (51.3, 62.1) with a 3-NN classifier?

   The 3NNs are (51,62), (51,63), (52,62). The labels of those are -1,1,1. So 3NN majority vote has label 1.

4. (Entropy) There is no fundamental reason that we measure entropy in bits. Let's assume that Na'vi (the alien from Avatar) makes memory units using water: each unit can be in one of three states of water (gas, liquid, solid). The information that each unit can hold is called a trit. What is the entropy in trits of a 5-sided die with probability (1/3, 1/9, 1/9, 1/9, 1/3)?

   sum_{i=1}^5 - pi log_3(pi) = -(2/3*(-1) + 3/9*(-2)) = 2/3+6/9=4/3 trits.

5. (Mutual information) The RDA Corporation has a prison with four cells. Without justification, you're about to be randomly thrown into a cell with equal probability. Cell 1 and cell 2 both have Toruks that eat prisoners. Cells 3 and 4 are safe. With sufficient bribe, the warden will answer your question "Will I be in cell 1?" What's the mutual information between the warden's answer and your encounter with the Toruks?

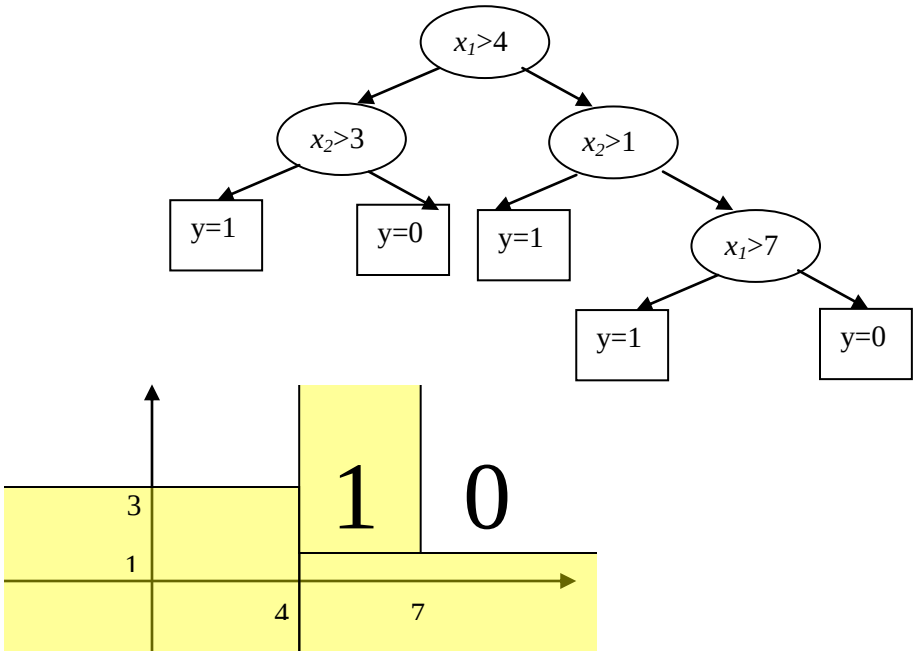
   P(W=y)=1/4, P(W=n)=3/4

   H(D|W=y)=0, H(D|W=n)=H(1/3, 2/3)

   H(D|W)=P(W=y)H(D|W=y) + P(W=n)H(D|W=n)=0+3/4H(1/3,2/3)

   H(D)=1. I(D;W)=H(D)-H(D|W)=1-3/4H(1/3,2/3)=0.3113

6. (Decision Tree) Consider the following decision tree on 2D continuous features. All left branches are "No" and right branches are "Yes". Draw the corresponding decision boundary in R^2.



7. (Linear algebra) Given a weight vector *w*, consider the line defined by $w'x$=3.14. Along this line, there is a point that is closest to the origin. How far is that point to the origin in Euclidean distance?

   3.14/||w||

8. (Support Vector Machines) Let w=(1, 2) and b=3. For the point x=(4,5), y=-1, what is the smallest slack value ξ for it to satisfy the margin constraint?

   The margin constraint is y(w'x+b)>=1-ξ. Plugging in the numbers, -(4+10+3)>=1-ξ, ξ>=18.

9. (Kernel) If K(x, x') is a valid kernel with induced feature representation ϕ(x), then G(x,x')=K(x,x')+3.14 is also a valid kernel. Write down the induced feature representation of kernel G using ϕ.

   This is a vector with one more dimension: [ϕ(x) sqrt(3.14)]

10. (Perceptron) With a linear threshold unit perceptron, implement the NAND function. That is, you should write down the weights w_0, w_A, w_B.

A	B	NAND
0	0	1
0	1	1
1	0	1
1	1	0

many possibilities, for example w0=1, wA= -0.75, wB= -0.75

11. (Gradient descent) Consider a linear perceptron $a = w_0+w_1*x_1+\ldots+w_d*x_d$. Given a single training point ($x_1=1$, $x_2=1$, …, $x_d=1$), $y=100$, compute the gradient at weight vector ($w_0=3$, $w_1=3$, …, $w_d=3$).

    E=0.5*(a-y)^2=0.5*(w0+w1+…wd-100)^2.
    grad = a d+1 dim vector with each element being (a-y)
    therefore, the gradient vector is a d+1 dim vector with elements (3d-97).

12. (Neural Network) You want to design a neural network with sigmoid units to predict a person's academic role from his webpage. Possible roles are "professor", "student", "staff". However, each person can take any number (from 0 to all 3) of these roles at the same time. Briefly describe (1) how you would represent the role label of a person in your training data, and (2) how to convert your neural network output (which will be real values because of the sigmoid units) to the roles.
    (1) the label of a person can be a vector of three Boolean values (PROF, STUD, STAF), each value can be 0 or 1 independent of the others. (2) The network needs three output units. Simply threshold each output at 0.5.

13. (Probability) 10% of the Na'vi's don't wear underwear, but they are too embarrassed to admit that. A surveyor wants to estimate that fraction and comes up with the following less-embarrassing scheme: Upon being asked "do you wear your underwear", a Na'vi would flip a fair coin outside the sight of the surveyor. If the coin ends up head, the Na'vi agrees to say "Yes"; otherwise the Na'vi agrees to answer the question truthfully. On a very large population, what fraction does the surveyor expect to hear the answer "Yes"?
    .5 + .5*(1-.1)=0.95.

14. (Probability) Given P(A|B)=0.4, P(B)=0.2, P(A)=0.5, compute P(B|A).
    P(B|A)=P(A|B)P(B)/P(A)=.4*.2/.5=.8/5=0.16.

15. (Bayesian network) Given the following network A➔B➔C where all variables are binary, write down the minimum number of (conditional) probabilities that define the CPTs.
    P(A), P(B|A), P(B|~A), P(C|B), P(C|~B)

16. (Estimating CPTs) You roll a 6-sided die 10 times and observe the following counts:
    side 1: 2, side 2: 3, side 3: 1, side 4: 0, side 5: 4, side 6: 0
    Use Laplace smoothing (i.e., add-1 smoothing), estimate the probability of each side.
    pi = (count(i)+1) / (total count + 6)
    p1=(2+1)/16=3/16, p2=4/16, p3=2/16, p4=1/16, p5=5/16, p6=1/16

17. (Speech) In one sentence, describe the McGurk effect.
    That we perceive some speech sounds through both acoustic and visual signal.

18. (Language Modeling) An n-gram language model computes the probability $P(w_n \mid w_1, w_2, \ldots, w_{n-1})$. How many parameters need to be estimated for a 5-gram language model?
Let V be the vocabulary size. Then for each (n-1)-gram history (the conditioning part), there are (V-1) parameters because the probability normalizes. There are V^(n-1) histories. So the number of parameters is (V-1)V^4. We also accept V^5.

19. (ESP) In one sentence, describe the foundation of the ESP game.
Inter-annotator agreement in lieu of true annotation accuracy control.

20. (Naïve Bayes) Consider a Naïve Bayes classifier with 100 feature dimensions. The label is binary with P(y=0)=P(y=1)=0.5. All features are binary, and have the same conditional probabilities:
$P(x_i=1 \mid y=0) = a$
$P(x_i=1 \mid y=1) = b$
for i=1, …, 100. The numbers a, b are in [0, 1]. Given an item with alternating feature values X= (x_1=1, x_2=0, x_3=1, …, x_{100}=0), compute P(y=1|X).
P(X|y=0)=prod P(xi|y=0)=a^50 (1-a)^50
P(X|y=1)=b^50 (1-b)^50
P(y=1|X)=P(X|y=1)P(y=1) / (P(X|y=1)P(y=1) + P(X|y=0)P(y=0)) = b^50 (1-b)^50 / (b^50 (1-b)^50 + a^50 (1-a)^50)